ML – Project 1

The steps that should be followed to do this project are:

1. Exploratory data analysis: [[1]](#footnote-1)

* Understand the dataset
* Analyze the features
* Which features are continuous
* Which features are categorical
* Are there any obvious relationships between features
* Look at the distribution of each feature

1. Feature processing:

* Clean the dataset
* Extract useful information
* Remove useless features and values
* Maybe **combine** features: NOTE: that for each method maybe a combination of features is better than other. We will begin with a fixed combination for all methods to follow this pipeline
* Find better **representations** of the features (I’m not sure what this means but it seems important to have a good feature transformation)

1. Implement the ML methods:

* Determine to see if it overfits or underfits (not very sure how to measure this but we’ll see)

1. Analyze each of the ML methods:

* Clearly describe the methods
* Understand what each model is exactly doing
* Which parameters could be improved (w\_initial, **lambda**, gamma, …) – REPRODUCIBILITY
* What is the impact of changing the parameters – BASELINES: clear experimental evidence
* Which datapoints are incorrectly classified and why

1. Chose the best model:

* Use cross-validation with **k-fold repetitions**
* Estimate the ‘**test error**’ - BASELINES

For the writing of the report we should use LaTeX and simply explain the upper steps (1-5) that we carried out.

**Exploratory data analysis:**

* Understand the dataset
* Analyze the features
* Which features are continuous
* Which features are categorical
* Are there any obvious relationships between features
* Look at the distribution of each feature

“Categorical variables contain a finite number of categories or distinct groups. Categorical data might not have a logical order. ... Continuous variables are numeric variables that have an infinite number of values between any two values.”

There are N=250.000 observations, and the feature matrix has 30 different features x0 , …, x29 . Then the ‘ids’ go from 100.000 until 349.999.

1. <https://es.wikipedia.org/wiki/An%C3%A1lisis_exploratorio_de_datos> [↑](#footnote-ref-1)